

# VEHICLE TRACKING USING ACOUSTIC AND VIDEO SENSORS \*

Aswin C Sankaranayanan, Qinfen Zheng, Rama Chellappa  
University of Maryland  
College Park, MD - 20770  
{aswch, qinfen, rama}@cfar.umd.edu

Volkan Cevher, James H. McClellan  
Georgia Institute of Technology  
Atlanta, GA 30332  
{gte460q, jim.mcclellan}@ece.gatech.edu

and

Gang Qian  
Arizona State University  
Tempe, AZ 85287  
gang.qian@asu.edu

## Abstract

In target tracking, fusing multi-modal sensor data under a power-performance trade-off is becoming increasingly important. Proper fusion of multiple modalities can help in achieving better tracking performance while decreasing the total power consumption. In this paper, we present a framework for tracking a target given joint acoustic and video observations from a co-located acoustic array and a video camera. We demonstrate on field data that tracking of the direction-of-arrival of a target improves significantly when the video information is incorporated at time instants when the acoustic signal-to-noise ratio is low.

familiar situation is the direction-of-arrival (DOA) tracking of multiple maneuvering targets in a two-dimensional plane using acoustic and video measurements. Video sensors can provide high resolution DOA estimates, but are constrained by their field of view. They also require relatively high power consumption. Acoustic sensors can be omni-directional and can track over the full  $360^\circ$  of DOA, and consume low power. However, acoustic sensors have difficulty distinguishing multiple targets in a convoy from a single target with harmonics.

## 1 Introduction

Detection, localization, and tracking by passive sensor arrays arise in many practical applications. A

We propose a sensor fusion framework based on particle filters [1] [2] to combine the detection and tracking results from a co-located acoustic array and video camera. Particle filter based trackers are used to recursively estimate the state probability density functions (pdf's) for the combined tracker. If the video controls the particle diversity at low signal-to-noise (SNR) region of the acoustics, the overall target tracking performance will be improved.

---

\*Prepared through collaborative participation in the Advanced Sensors Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-02-0008.

## 2 Sensor Models

The acoustic sensor array consists of a number of sensors uniformly distributed along a circle. The camera is assumed to be located in the center of the circle. Further, complete knowledge of sensor calibration, such as camera focal length and acoustic node location is assumed.

The acoustic system acquires the target DOA under the assumptions of narrow-band target frequencies, constant velocity, motion and target in the far-field of the sensor array. The acoustic state vector for the target at time instant  $t$  is given by  $\mathbf{x}_{a,t} = \{\theta_t, q_t, \phi_t\}$ , where  $\theta_t, q_t, \phi_t$  are the DOA, logarithmic ratio of the target speed to its range, and the vehicle heading direction, respectively. A particle filter was used to recursively estimate the state of the target. Details can be found in [3–5].

The video state vector for the target at time  $t$  is given by  $\mathbf{x}_{v,t} = \{\alpha_t, \mathbf{u}_t, \mathbf{v}_t\}$ , where  $u_t, v_t$  are the coordinates of the centroid of the tracked object on the image plane and  $\alpha_t$  accounts for affine distortions on the target size and orientation. The video tracker assumes an adaptive velocity model for the target’s motion, estimating a velocity based on first order linear approximations on appearance differences between the incoming observation (image) and the previous particle configuration. Details can be found in [6]. The video tracker also has a built-in self-evaluation that is used to detect tracking failure. When such failure occurs, the track is terminated and re-initialized using a motion based detector. This evaluation uses information gathered from multiple cues such as appearance, track stability, aspect ratio of the target size to detect tracking failure. Details of this can be found in [8].

The interaction between the acoustics and video sensor depends on coordinate transformations between the state-spaces. Given the video state parameters, it is possible to transform them to the acoustic state space as follows:

$$\begin{aligned} \tan \theta_t &= \frac{u_t}{f} \\ q_t &= \frac{\sqrt{(\Delta u)^2 + (\Delta v)^2} \cos \theta}{T \sqrt{u^2 + v^2} \cos \phi} \end{aligned} \quad (1)$$

where  $f$  is the focal length of the camera and  $T$  is the sampling period of the video.

## 3 Fusion Algorithm

In this section, a fusion strategy is given where the video helps the acoustic tracker when the target acoustic SNR is low in order to obtain a better overall track [7]. The key step in the algorithm is to use video’s high target resolving capability to propose particles for the acoustic tracker. Under the assumption that the video tracker has not lost track of the target (and the self-evaluation method detects tracking failure and terminate tracks), the particles generated by the video are bound to be in close proximity of the true target.

There are fundamental differences in the working of acoustic and video trackers, partially due to the differences in acoustic and video data. These play an important role in determining the nature of the joint tracker. Some of these are listed below:

- The sampling rate for video data (for the dataset used) is 30 frames per second and the tracker estimates density functions for target DOA for each frame, giving us 30 DOA estimates per second.
- Acoustic information is captured at 1024 samples per second (for the dataset used) and the acoustic tracker groups the samples in frames of duration  $T_0$  seconds to estimate the target DOA. Further, given a frame of acoustic data from time  $t_1$  to time  $t_1 + T_0$ , the acoustic tracker estimates target DOA for the beginning of the time interval,  $t_1$ .
- Acoustic signals also suffer from a significant propagation delay that cannot be estimated as we have no information about the range of the target. Video data does not suffer from a similar propagation delay.

Given the differences in the rate of generation of DOA, sensor observations are merged periodically, say every  $T$  seconds. For our experiments,  $T_0$ , the length of the acoustic frame is set equal to  $T$ , the delay between two instants of data fusion and the value chosen for  $T$  (and hence,  $T_0$ ) is 1 second. Given the availability of both acoustic and video data for an interval of time, the video tracker participates in data fusion once every 30 of its estimation cycles while the acoustic tracker participates for every acoustic frame. Figure 1 illustrates the basic idea behind the proposed tracker.

Let  $\{\mathbf{x}_{v,t}^{(i)}, \mathbf{w}_{v,t}^{(i)}\}$  and  $\{\mathbf{x}_{a,t}^{(i)}, \mathbf{w}_{a,t}^{(i)}\}$  denote the target motion parameter samples and their associated weights at time  $t$  obtained through video and

acoustic tracking, respectively. Note that the acoustic frame contains data samples from time interval  $(t, t + T_0)$ . The major steps of the fusion algorithm are listed below:

- i. Video proposes  $\beta N$  particles ( $0 < \beta < 1$ ) by using the previously estimated target states. Transformation equations are used to convert the video motion parameters to acoustic parameters.
- ii. The acoustic tracker proposes another set of  $(1 - \beta)N$  based on its previous estimate and linear motion model.
- iii. These  $N$  particles are updated using the acoustic data likelihood probabilities. The acoustic tracker's output forms the joint tracker's estimate.

The parameter  $\beta$  depends on the confidence measures of the individual trackers. For experiments conducted to test the proposed algorithm,  $\beta = .5$  was used. Hence, half of the particles for the acoustic tracker are proposed by the video tracker. The value of  $T$  used was 1 second.

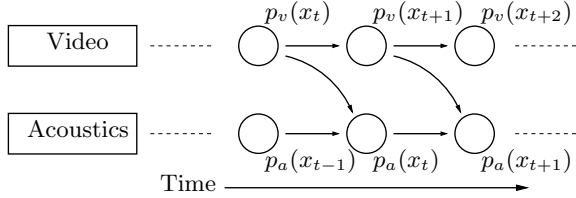


Figure 1: Previous video motion estimates are also used to propose particles for the acoustic tracker. The acoustic data observed is used to update these states with weights.

## 4 Results

The proposed fusion algorithm was tested with field data collected at Aberdeen Proving Grounds. The target tracked was an M60 tank, with the track of the target shown in Fig. 2. The acoustic array has four microphones located on a circle of radius equal to .5m. The video data consists of IR images with pixel resolution  $720 \times 480$ . The field of view is approximately 18 degrees, 11 degrees to the left of the z-axis (see Fig. 3).<sup>1</sup>

<sup>1</sup>The reference axes used for DOA definition by the video and acoustic tracker were not aligned because of the structure of the sensor array. Estimates of DOA were aligned prior to fusion.

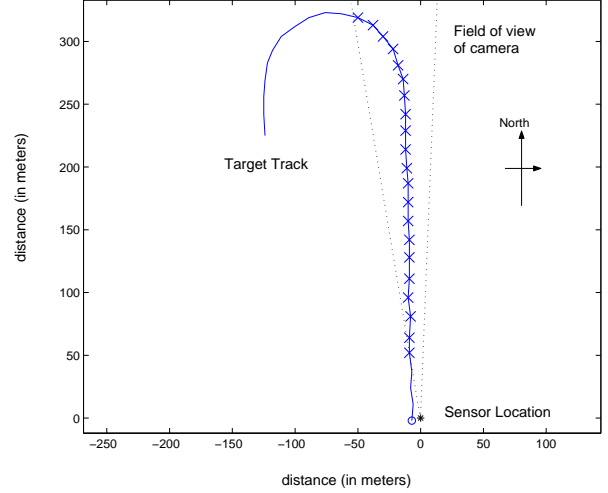


Figure 2: Track of the target with origin defined at the sensor location. Initial position of the target is circled and the portion of the track that had video coverage is starred. Experiments conducted had target initial and final locations as shown.

The orientation of the field of view and the relative positions of the acoustic sensors are shown in Fig. 3. In reality the acoustic sensor array and the

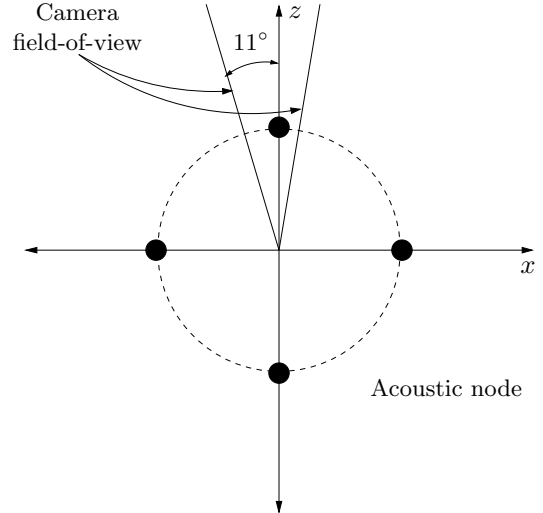


Figure 3: The acoustic microphones (black dots) are omnidirectional and are co-located with the camera as shown.

IR camera used were not co-located (the centers were displaced by a few meters). The underlying theory doesn't take into account this factor. This is not significant when the target is far away from the sensor array/IR camera. Note that the acoustic tracker also

works under the far-field assumption on the target. This, however, leads to errors in DOA estimation at the initial time instants of tracking when the target is close to the sensor (see Fig. 2). The proposed tracker assumes knowledge of parameters like the focal length of the IR camera. The focal length was found by manual calibration. The track of the target on the ground-plane is shown in Fig. 2.

Two scenarios were considered in testing the proposed algorithm. In the first scenario, video information is used whenever the target is in the field of view of the IR camera. The DOA estimates for this test are shown in Fig. 4.

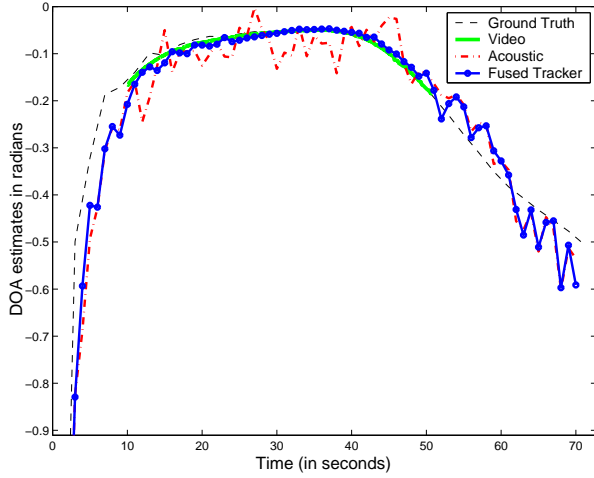


Figure 4: Plot shows improved tracking using proposed algorithm.

The following can be observed from the figure:

- Estimates of DOA from the video tracker deviate from the ground truth by as much as 0.02 radians (approximately 1 degree) at initial time instants. This is caused because of errors in measurement of the focal length and GPS magnified by the fact that the target is close to the sensor. Note that the video tracker uses the centroid of the bounding box of the target for deriving DOA estimates. When the target is close to the sensor, the centroid might not be representative of the location of the GPS on the target.
- The proposed joint acoustic video tracker (or the fused tracker) has a smaller variance (about the ground truth) than the acoustic-only tracker. This is because the particles are proposed by the video tracker whose DOA resolution is better than the acoustic-only tracker.

The DOAs of the particles proposed by the video tracker at each time instant are shown in Fig. 5. Note that most of the particles proposed are in the neighborhood of the ground truth. Such a set of particles with low variance (and in a close proximity to the truth) in turn leads to better DOA estimation by the fused tracker. Figure 6 shows the DOA values of the particles proposed by the acoustic tracker for the segment of the track when the SNR is low (due to propagation losses).

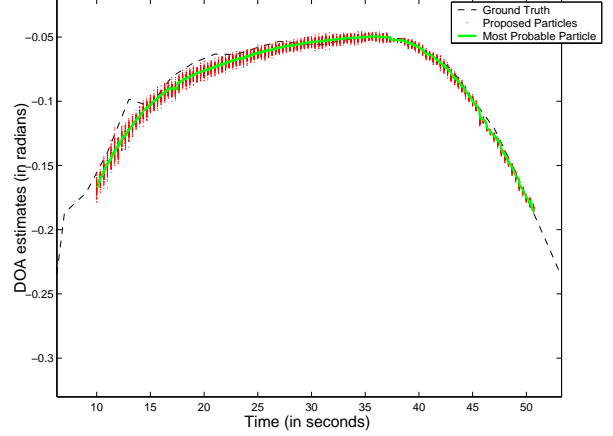


Figure 5: Plot showing the DOAs of the particles proposed by the video tracker.

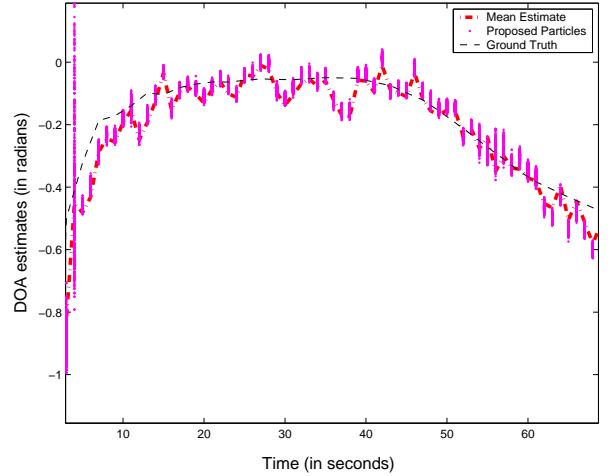


Figure 6: Plot showing the DOAs of the particles proposed by the acoustic tracker.

The second scenario considered for testing the fusion algorithm was to have the video camera (and hence the video tracker) powered for sections of the track when the acoustic tracker receives signals with

low SNR. In our experiment, such a scenario was *simulated* by using video information for selective time segments. During these segments, the acoustic tracker used the particles proposed by the video tracker and using its own proposal scheme during other time segments. Figure 7 shows the performance of the proposed tracker under such a scenario. It was noted the performance improved significantly but only in a short period of time (about 10 seconds) after the video was turned off. Such a scenario can be very useful in getting better acoustic performance for low SNR signals. The power required to sustain the video camera is cut by 75% when compared to the first scenario ( Fig. 4).

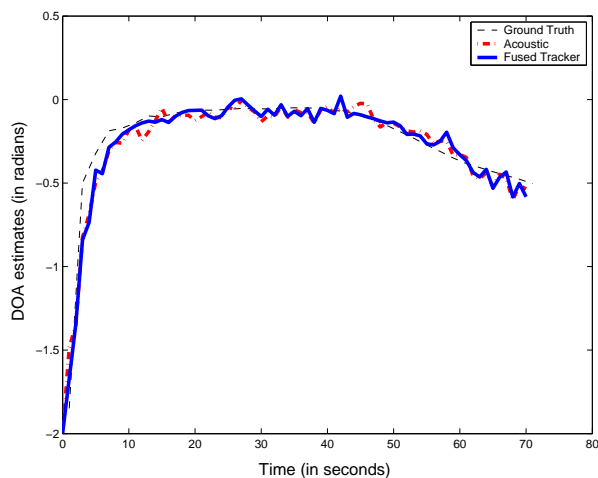


Figure 7: Plot showing performance of the trackers when video information was selectively used over different segments of time. In this experiment, the video information was used at time intervals [11, 15] and [44, 49]. Note how the variance of the estimates of the fused tracker (about the ground truth) decreases in the vicinity of [11, 15] and [44, 49].

The time evolution of the results of the first scenario are shown in Figures (8-12).

## 5 Conclusion

A framework for sensor fusion is proposed in this paper. When using fusion, the tracker takes advantage of the omni-directional sensing field of the acoustic sensor as well as the estimation accuracy of the video camera. In this framework, particles proposed by a video tracker guide the prediction of the acoustic tracker. Improved DOA estimation is demonstrated by the proposed fusion framework.

## References

- [1] A. Doucet, N. d. Freitas, and N. Gordon, "Sequential Monte Carlo Methods in Practice." Springer-Verlag, New York, 2001.
- [2] J. S. Liu and R. Chen, "Sequential Monte Carlo Methods for Dynamic Systems," *Journal of the American Statistical Association*, vol. 93, pp 1031-1041, 1998.
- [3] M. Orton and W. Fitzgerald, "A Bayesian approach to tracking multiple targets using sensor arrays and particle filters," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 216-223, 2002.
- [4] Y. Zhou, P.C. Yip, and H. Leung, "Tracking the direction-of-arrival of multiple moving targets by passive arrays: Algorithm," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2655-2666, 1999.
- [5] V. Cevher and J. H. McClellan, "General direction-of-arrival tracking with acoustic nodes," to appear *IEEE Trans. Signal Processing*.
- [6] S. Zhou, R. Chellappa, and B. Moghaddam, "Adaptive visual tracking and recognition using particle filters," to appear *IEEE Trans. Image Processing*, Nov 2004.
- [7] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle detection and tracking using acoustic and video sensors," *ICASSP 2004*, Montreal, CA, 17-21 May 2004.
- [8] H. Wu and Q. Zheng "Self-evaluation for video tracking systems," to appear *Army Science Conference*, 2004

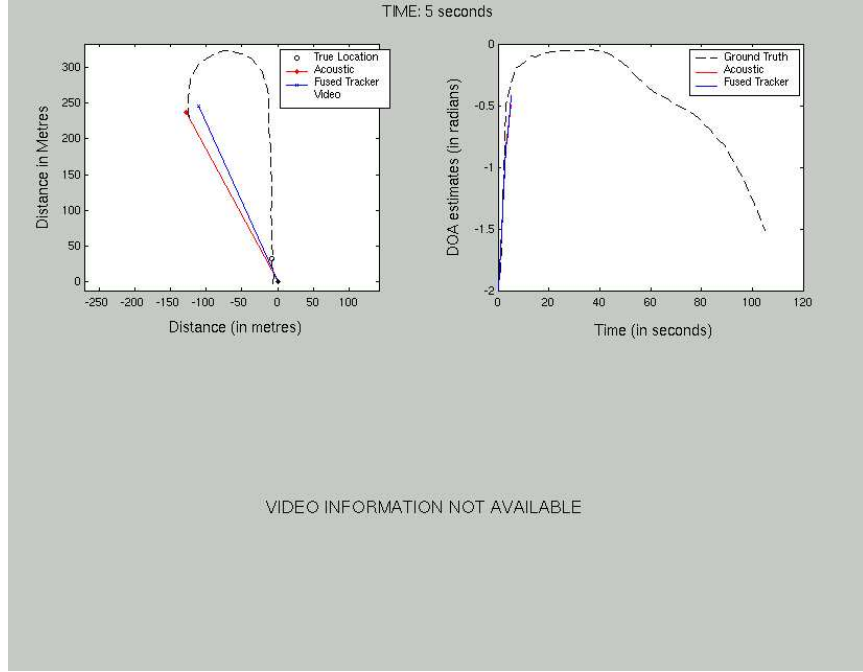


Figure 8: A screen-shot showing DOA plots and position information of acoustic and fused tracker for time  $t = 5$  seconds. The experiment was started at time  $t = 0$  seconds. No video information was available as target was not in the field of view of the camera.

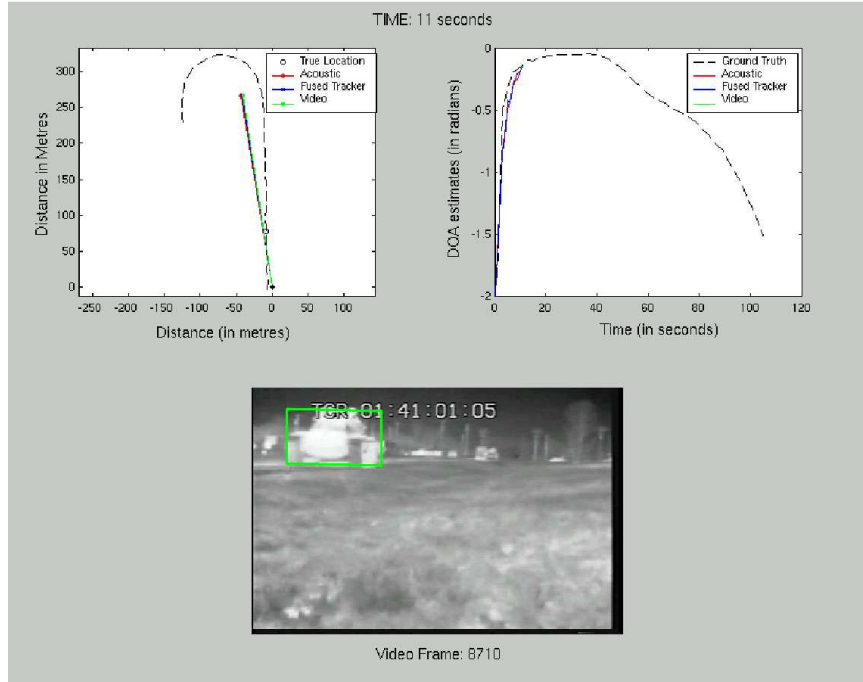


Figure 9: Screen-shot showing DOA plots and position information of acoustic, video and fused tracker for time  $t = 11$  seconds. Inset is the image showing the estimate of the video tracker.

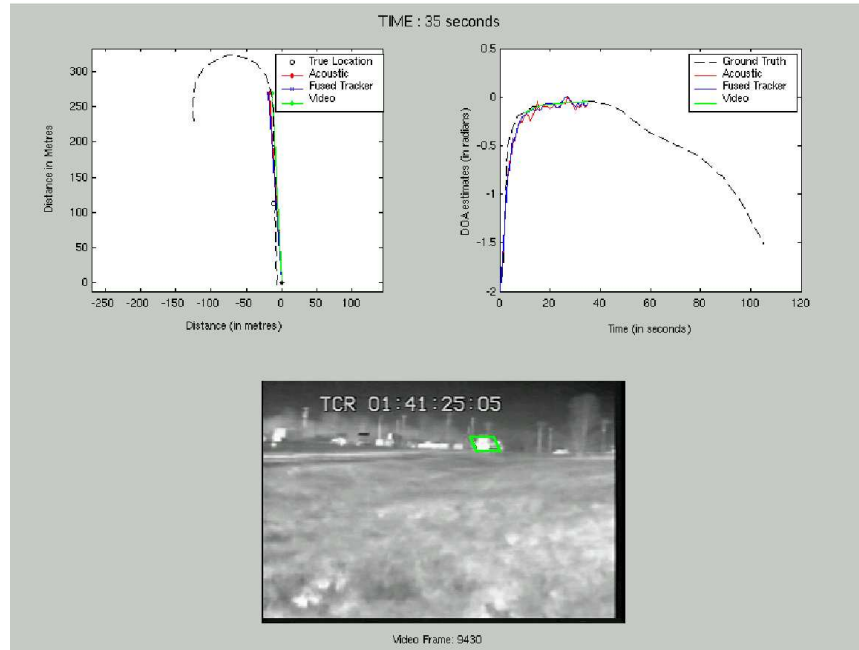


Figure 10: Screen-shot showing DOA plots and position information of acoustic, video and fused tracker for time  $t = 35$  seconds. Inset is the image showing the estimate of the video tracker.

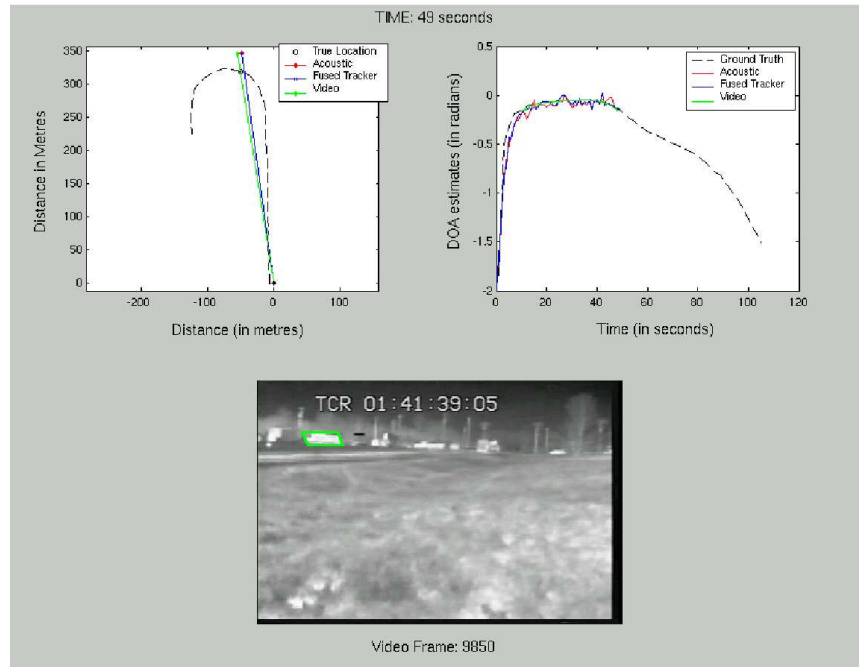


Figure 11: Screen-shot showing DOA plots and position information of acoustic, video and fused tracker for time  $t = 49$  seconds. Inset is the image showing the estimate of the video tracker.

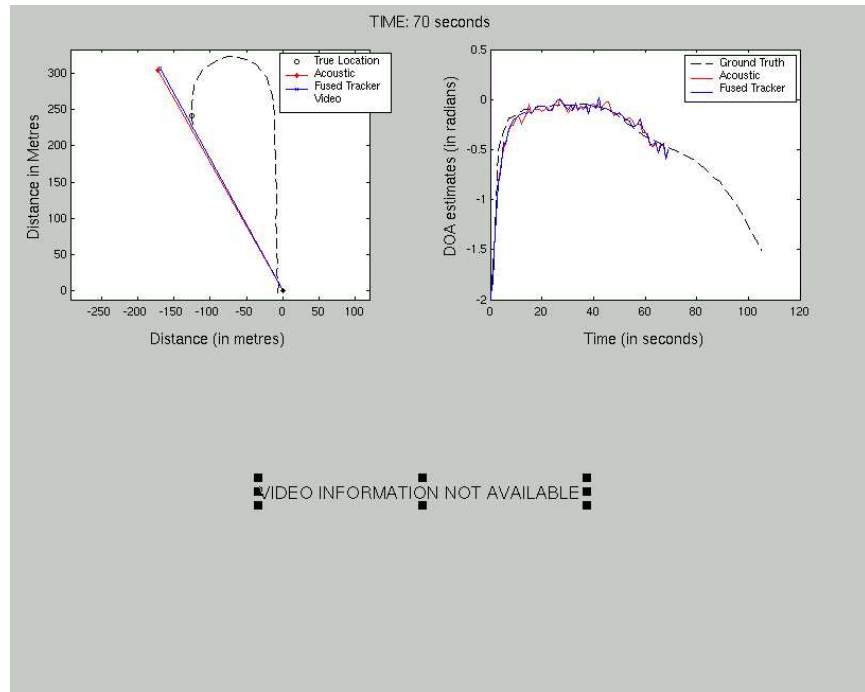


Figure 12: Screen-shot showing DOA plots and position information of acoustic and fused tracker for time  $t = 70$  seconds. No video information was available as target was not in the field of view of the camera.